

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平7-219929

(43) 公開日 平成7年(1995)8月18日

(51) Int.Cl.⁶

G 0 6 F 17/18

G 0 5 B 19/418

識別記号

庁内整理番号

F I

技術表示箇所

7531-3H

G 0 6 F 15/ 36

Z

G 0 5 B 15/ 02

S

審査請求 未請求 請求項の数18 O L (全 32 頁)

(21) 出願番号 特願平6-8871

(22) 出願日 平成6年(1994)1月28日

(71) 出願人 000006013

三菱電機株式会社

東京都千代田区丸の内二丁目2番3号

(72) 発明者 上田 太郎

横浜市戸塚区川上町87番地1 三菱電機東

部コンピュータシステム株式会社横浜シス

テムセンター内

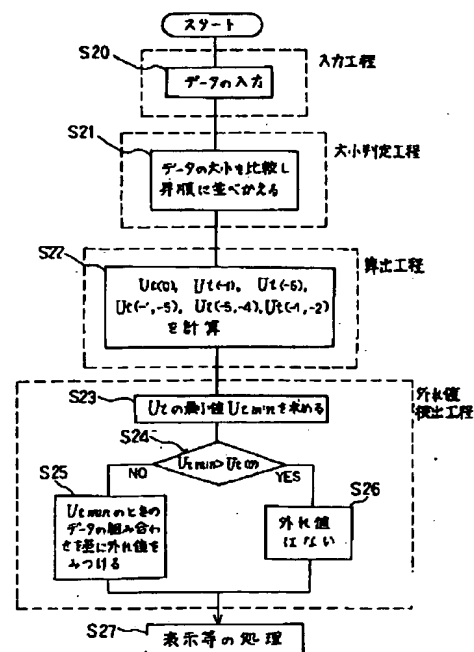
(74) 代理人 弁理士 高田 守

(54) 【発明の名称】 外れ値検出方法及びデータ処理装置

(57) 【要約】

【目的】 得られたデータ（収量、反応量）の外れ値を求める。

【構成】 入力工程でデータを入力し、大小判定工程で昇順に並べる。次に、算出工程で一番小さい値を除いた検出統計量を計算する。一番大きい値を除いた検出統計量を計算する。一番小さい値と一番大きい値を除いた検出統計量を計算する。以下同様にして考えられる組み合わせのデータから検出統計量を計算する。またデータを除かないときの検出統計量も計算する。外れ値検出工程で、以上の検出統計量が最小となるデータの組み合わせを見つけ、その除いた値を外れ値とする。データを除かないときの検出統計量が最小となつたとすれば外れ値は存在しない。



1

【特許請求の範囲】

【請求項1】 以下の工程を有する外れ値検出方法

(a) N 個 ($N \geq 3$) の値を入力する入力工程、(b) 上記入力工程により入力した N 個の値の大小関係を判定する大小判定工程、(c) 上記大小判定工程により判定された大小関係に基づき、 N 個の値の組み合わせ及び外れ値の候補を除いた N 個未満の値の組み合わせを求め、求めた組み合わせに対して所定の計算式を用いて検出統計量を算出する算出工程、(d) 上記算出工程により算出された検出統計量に基づいて、外れ値を検出する外れ値検出工程。

【請求項2】 上記算出工程は、 s 個以内の外れ値を検出する場合、大小判定工程により判定された大小関係上連続する n 個 ($n = N - s$) 以上の値の複数の組み合わせを用いて検出統計量を算出することを特徴とする請求項1記載の外れ値検出方法。

【請求項3】 上記外れ値検出工程は、 N 個未満の値の組み合わせから求めた検出統計量の中で最小のものを選択する最小値選択工程と、選択された最小値が N 個の値の組み合わせから求めた検出統計量よりも小さい場合に、その選択された最小値を算出した組み合わせに含まれていなかった値を外れ値とする外れ値判定工程を備えたことを特徴とする請求項1又は2記載の外れ値検出方法。

【請求項4】 上記計算式は、外れ値の候補が除かれると小さくなる傾向にある第1の項目と、外れ値の候補が除かれると大きくなる第2の項目とを有し、上記算出工程は、第1と第2の項目の値を算出し両者の和により検出統計量を求めることを特徴とする請求項1、2又は3記載の外れ値検出方法。

【請求項5】 上記計算式は、更に、第1と第2の項目以外に、第1と第2の項目を補正する補正項を有し、上記算出工程は、第1と第2と第3の項目の値を算出し、3者の和により検出統計量を求めることを特徴とする請求項4記載の外れ値検出方法。

【請求項6】 上記第1の項目は、検出統計量を求める N 個未満の値の分散を用いていることを特徴とする請求項4又は5記載の外れ値検出方法。

【請求項7】 上記第2の項目は、検出統計量を求める場合の外れ値の候補の個数を用いていることを特徴とする請求項4又は5記載の外れ値検出方法。

【請求項8】 上記第2の項目は、外れ値の候補の個数に対して所定の係数を乗算したものをを用いることを特徴とする請求項7記載の外れ値検出方法。

【請求項9】 上記計算式は、検出統計量を求める N 個未満の値の分散と分散に対する係数を有しており、上記算出工程は、分散と係数の乗算により検出統計量を求めることを特徴とする1、2又は3記載の外れ値検出方法。

【請求項10】 上記計算式は、回帰分析の変数選択基準を基礎にして作成されることを特徴とする請求項1～

2

8又は9記載の外れ値検出方法。

【請求項11】 上記外れ値検出方法は、更に、入力工程と大小判定工程の間に、入力した値を加工する加工工程を備えたことを特徴とする請求項1記載の外れ値検出方法。

【請求項12】 上記加工工程は、入力工程により入力された時間に依存する値を時間に依存しない値に加工することを特徴とする請求項11記載の外れ値検出方法。

【請求項13】 上記加工工程は、入力工程により入力された値からデコ比を計算することを特徴とする請求項11記載の外れ値検出方法。

【請求項14】 上記加工工程は、入力工程により入力された値から回帰分析モデルのデータを計算することを特徴とする請求項11記載の外れ値検出方法。

【請求項15】 上記加工工程は、入力工程により入力された値から正準相関分析モデルのデータを計算することを特徴とする請求項11記載の外れ値検出方法。

【請求項16】 上記加工工程は、入力工程により入力された値が複数のグループに分類されていて複数の要因により判別分析を行う場合に、各グループの判別関数値を計算することを特徴とする請求項11記載の外れ値検出方法。

【請求項17】 上記請求項1～15又は16記載の外れ値検出方法を実行して外れ値を検出する外れ値検出手段と、 N 個の値を計測して外れ値検出手段に入力する計測手段と、外れ値検出手段により検出された外れ値を知らせる出力手段を備えたデータ処理装置。

【請求項18】 上記データ処理装置は、更に、外れ値検出手段により検出された外れ値を除いた残りの値を用いて所定の処理を実行するデータ処理手段を備えたことを特徴とする請求項17記載のデータ処理装置。

【発明の詳細な説明】

【0001】

【産業上の利用分野】 この発明は生産工程、品質管理、研究開発、品質改良などにおけるデータの外れ値を検出する方法及びその方法を利用した装置に関するものである。

【0002】

【従来の技術】 例えば、製品の性能バラ付きを測定する場合、あるいは、電力メータや水道メータ等の検針を行う場合、更には実験データを測定する場合に、得られたデータの中に規格外れの性能を示すデータや、異常な測定値を示すデータが存在する。このように、規格外れのデータや、異常値は測定環境や測定装置自身から生ずる不適切なデータであることが多い。このような不適切なデータを、ここでは以下外れ値と呼ぶことにする。外れ値は、本来測定されるべき値ではないため、前述したような各種データから外れ値を検出し、取り除く手法が従来から考えられてきている。データから外れ値を検出する方法は、従来から統計手法に基づくものがある。外れ

3

値とは極端に大きなあるいは小さい値をとるデータのことである。例えば、5.71、6.57、7.29、8.06、10.00、15.00を考える。プロットすると図12のようになる。図12を見ると15.00は外れ値のようである。統計手法では外れ値が1個として、1個の時の外れ値を検定する計算式を用いる。2個の時は2個用の計算式を用いる。統計的検定であるから予め危険率（有意水準）を決めておく必要がある。危険率としては伝統的に5%あるいは1%を用いている。危険率5%とは、統計的検定により外れ値と判断を下す時誤る確率が5%であることを示す。計算式に対応した5%あるいは1%の教表があり、実データで計算した値と教表とを比較して大ならば外れ値とする。ただし危険率としては、5%ある。または1%あるということになる。

【0003】このように従来の統計手法では外れ値の個数が1個の時、2個の時、3個の時によって計算法が異なったり、危険率（有意水準とも呼ばれる）の違い（5%、1%等）により結論が異なる（5%の時外れ値と結論しても1%の時外れ値とはいえない等）問題点がある。また大きな値の外れ値、小さな値の外れ値により計算法が異なる。統計手法であるから5%、1%の教表も必要である。

【0004】具体的には、図13及び図14を用いて説明する。図13で1は情報処理装置、2はコンピュータ（FDD付）、3はディスプレイ・ユニット、4はプリンタ、5はキーボード、6はフロッピーディスクである。プログラム・ルーチンが記憶されたフロッピーディスク6をコンピュータ（FDD付）2に挿入し、オペレ*

$$T_i = (x_i - \bar{x}) / S$$

$$\text{ただし } S = \sum (x_i - \bar{x}) / (n-1), \quad n: \text{ サンプル数}$$

【0009】15.00が外れ値と考えられるので、 T_i の最大値 $\max T_i$ ($i=1, 2, \dots, n$) を求めて教表に載っている値と比較する。

$$\max T_i = 1.841$$

となった。Grubbsの教表、表1を見るとサンプル数 $n=6$ の時、かつ、危険率5%の時1.82、サンプル数 $n=6$ 、かつ、危険率1%の時1.94である。よ

$$\max T_i > 1.82, \max T_i < 1.94$$

である。従って、15.00は危険率5%で外れ値といえる。危険率1%では外れ値といえない。このように危険率の違いにより結論が異なってくる。

【0010】

【表1】

4

*ーション・ソフトを駆動して、情報処理装置1をスタートさせる。フロッピーディスク6からプログラム・ルーチンがロードされ入力待状態となる。

【0005】図14は従来例の説明のためのフローチャートである。ステップ1は、キーボード5からデータを連続的に入力する段階である。ステップ2では、外れ値の個数を入力し、1ならば、ステップ3、ステップ4で、小さい値又は大きい値を外れ値とした統計量をそれぞれ求める。なぜ別々に求めるかと言えば、小さい値と大きい値では計算方式が異なるからである。また、外れ値の個数が2個の場合は、ステップ5からステップ7で、小さい値を2個外れ値とした場合、小さい値と大きい値を1個ずつ外れ値とした場合、大きい値を2個外れ値とした場合にそれぞれ別の計算方式で統計量を求める。

【0006】ステップ8では、教表を見て上記ステップで求めた計算値と教表にある有意点の大小比較をする。ステップ9では、有意点より計算値の方が大きい場合外れ値と認識する。ステップ10は、計算値の方が小さい場合外れ値としない。ステップ11では、結果の表示等をする。図14は危険率（有意水準）が5%の場合であるが、1%の場合なら1%の教表が必要となる。

【0007】従来の統計手法で外れ値を検出する例を示す。データとして、

5.71、6.57、7.29、8.06、10.00、15.00

とする。Grubbs検定量の式、教1を用いる。

【0008】

【教1】

Grubbsの教表の一部

n	5%	1%
4	1.46	1.49
5	1.67	1.75
6	1.82	1.94
7	1.94	2.10

【0011】(Vic Barnett, Toby Lewis (1978): 「Outliers in Statistical Data」, John Wiley & Sons. p. 298から一部引用)

【0012】次に、マスク効果の例をあげる。データとして5.71、6.57、7.29、8.06、14.80、15.00とする。このデータでは $\max T_i = 1.29$ となる。 $\max T_i < 1.82$ である。従って外れ値はないことになる。これはマスク効果といって、

5

上のように外れ値の候補が14.80と15.00の2つある場合、従来方式では外れ値を1つとして検定すると必ずしも外れ値を検出しない例である。

【0013】

【発明が解決しようとする課題】以上説明したように、従来のものでは外れ値の個数により計算方式が異なる。また、外れ値の性格（大きい方の外れ値か小さい方の外れ値か）により計算方式が異なる（Vlc Barnett, Toby Lewis (1978): 「Outliers in Statistical Data」, John Wiley & Sonsには40種以上の計算式が載っている）という問題点があった。また、計算値と致表の大小比較が必要である。また、危険率の違い（5%、1%等）により結論が異なるという問題点があった。また、マスク効果といって例えば外れ値の候補が2つある場合、従来方式では外れ値を1つとして検定すると必ずしも外れ値を検出しないという問題点があった。

【0014】この発明は、以上のような問題点を解決するためになされたものであり、従来のような致表を用いることなく、また、外れ値の個数や外れ値の性格により計算方式を変える必要がない外れ値検出方法を得ることを目的とする。また、マスク効果を回避することができる外れ値検出方法を得ることを目的とする。また、外れ値を検出する場合にできるだけ計算過程が簡単で、且つ、計算量も少なく済む外れ値検出方法を得ることを目的とする。更には、これらの外れ値検出方法を利用したデータ処理装置を提供することを目的とする。

【0015】

【課題を解決するための手段】この発明に係る外れ値検出方法は、以下の工程を有する。

(a) N個 ($N \geq 3$) の値を入力する入力工程、(b) 上記入力工程により入力したN個の値の大小関係を判定する大小判定工程、(c) 上記大小判定工程により判定された大小関係に基づき、N個の値の組み合わせ及び外れ値の候補を除いたN個未満の値の組み合わせを求め、求めた組み合わせに対して所定の計算式を用いて検出統計量を算出する算出工程、(d) 上記算出工程により算出された検出統計量に基づいて、外れ値を検出する外れ値検出工程。

【0016】上記算出工程は、s個以内の外れ値を検出する場合、大小判定工程により判定された大小関係上連続するn個 ($n = N - s$) 以上の値の組み合わせを複数作成し、これらの組み合わせを用いて検出統計量を算出することを特徴とする。

【0017】上記外れ値検出工程は、N個未満の値の組み合わせから求めた検出統計量の中で最小のものを選択する最小値選択工程と、選択された最小値がN個の値の組み合わせから求めた検出統計量よりも小さい場合に、その選択された最小値を算出した組み合わせに含まれていなか

6

った値を外れ値とする外れ値判定工程を備えたことを特徴とする。

【0018】上記計算式は、外れ値の候補が除かれると小さくなる傾向にある第1の項目と、外れ値の候補が除かれると大きくなる第2の項目とを有し、上記算出工程は、第1と第2の項目の値を算出し両者の和により検出統計量を求めることを特徴とする。

【0019】上記計算式は、更に、第1と第2の項目以外に、第1と第2の項目を補正する補正項を有し、上記算出工程は、第1と第2と第3の項目の値を算出し、3者の和により検出統計量を求めることを特徴とする。

【0020】上記第1の項目は、検出統計量を求めるN個未満の値の分散を用いていることを特徴とする。

【0021】上記第2の項目は、検出統計量を求める場合の外れ値の候補の個数を用いていることを特徴とする。

【0022】上記第2の項目は、外れ値の候補の個数に対して所定の係数を乗算したものをを用いることを特徴とする。

【0023】上記計算式は、検出統計量を求めるN個未満の値の分散と分散に対する係数を有しており、上記算出工程は、分散と係数の乗算により検出統計量を求めることを特徴とする。

【0024】上記計算式は、回帰分析の変数選択基準を基礎にして作成されることを特徴とする。

【0025】上記外れ値検出方法は、更に、入力工程と大小判定工程の間に、入力した値を加工する加工工程を備えたことを特徴とする。

【0026】上記加工工程は、入力工程により入力された時間に依存する値を時間に依存しない値に加工することを特徴とする。

【0027】上記加工工程は、入力工程により入力された値からデコ比を計算することを特徴とする。

【0028】上記加工工程は、入力工程により入力された値から回帰分析モデルのデータを計算することを特徴とする。

【0029】上記加工工程は、入力工程により入力された値から正相関分析モデルのデータを計算することを特徴とする。

【0030】上記加工工程は、入力工程により入力された値が複数のグループに分類されていて複数の要因により判別分析を行う場合に、各グループの判別関数値を計算することを特徴とする。

【0031】また、この発明に係るデータ処理装置は、外れ値検出方法を実行して外れ値を検出する外れ値検出手段と、N個の値を計測して外れ値検出手段に入力する計測手段と、外れ値検出手段により検出された外れ値を知らせる出力手段を備える。

【0032】上記データ処理装置は、更に、外れ値検出手段により検出された外れ値を除いた残りの値を用いて

所定の処理を実行するデータ処理手段を備えたことを特徴とする。

【0033】

【作用】第1の発明においては、入力工程により、N個の値が入力されると、大小判定工程により値の大小関係を判定し、大きい方の値又は小さい方の値のいくつかを外れ値の候補とする。算出工程は、まずN個の値の組み合わせ及び外れ値の候補を除いたN個未満の値の組み合わせを求め、次に求めた組み合わせそれぞれに対して所定の計算式を用いて検出統計量を算出する。外れ値検出工程は、算出された検出統計量に基づいて外れ値を検出する。

【0034】第2の発明における算出工程は、大小判定工程により判定された値の大小に基づき、n個 ($n = N - s$) 以上の連続する値の組み合わせを用いて、所定の計算式により検出統計量を計算する。例えば、入力工程により5個 ($N = 5$) が入力され、最大2個 ($s = 2$) の外れ値を検出しようとする場合、大きい方から3個の入力値を用いて1つの組み合わせを作成する。また、大きい方から4個の入力値を用いて別な組み合わせを作成する。また、最大値と最小値を除いた中間の値3個を用いて1つの組み合わせを作成する。また、小さい方の入力値3個及び小さい方の入力値4個を用いてそれぞれ組み合わせを作成する。

【0035】第3の発明における外れ値検出工程は、まず、算出工程により算出されたN個未満の値の組み合わせの検出統計量の中で最小のものを選択する。次に、N個の値の組み合わせから求めた検出統計量と、選択された最小値を比較し最小値の方が小さい場合、その選択された最小値を算出した組み合わせに含まれていなかった値を外れ値とする。また、最小値の方が大きい場合、外れ値は無しと判定する。

【0036】第4の発明における計算式は、外れ値の候補が除かれると小さくなる傾向にある第1の項目と、外れ値の候補が除かれる時に大きくなる第2の項目を有し、両者の和により検出統計量を求める。この計算式により、外れ値がある場合最も外れた値が除かれると検出統計量が最小となる。

【0037】第5の発明における計算式は、上記第1と第2の項目に加えて、第3の項目を持つ。この第3の項目は、上記第1と第2の項目を補正する補正項目である。第1項目、第2項目、第3項目を加算して検出統計量を求める。

【0038】第6の発明における計算式は、第1項目に検出統計量を求めるN個未満の値の分散を含んでいる。従って、最も外れた値が除かれると分散の値が小さくなり、第1の項目の値が小さくなる。

【0039】第7の発明における計算式は、第2項目に検出統計量を求める場合の外れ値の候補の個数を含んでいる。従って、外れ値の数を多く検出しようとする、

第2の項目の値が大きくなる。

【0040】第8の発明における計算式は、第2の項目に外れ値の候補の個数に対して所定の係数を乗算したものをを用いる。

【0041】第9の発明における計算式は、検出統計量を求めるN個未満の値の分散に係数を乗算して検出統計量を求める。

【0042】第10の発明における計算式は、回帰分析の変数選択基準を基礎にして検出統計量を求める計算式を作成する。

【0043】第11の発明においては、加工工程により、入力工程により入力された値を、検出統計量を求めることができるデータに変換することができるため、様々な種類のデータを入力することができる。

【0044】第12の発明においては、入力工程により入力された値から例えば時間に比例して増加、あるいは、減少する傾向を補正して時間に依存しない値に加工する。そして、補正された値から検出統計量を算出し、外れ値を求めることができる。

【0045】第13の発明においては、1つのサンプルに複数の特性値がある場合に、デコ比の対角要素を計算し、計算された値を基に検出統計量を求め外れ値を求める。

【0046】第14の発明においては、入力された値が回帰分析の手法を適用できる場合、回帰分析の残差を計算し、計算された値を基に検出統計量を求め外れ値を求める。

【0047】第15の発明においては、入力された値が正準相関分析モデルのデータの場合、正準相関分析を行い合成変数関数値を2個求め、これより合成変数関数値を求め、合成変数関数値からデコ比を計算しデコ比の計算された値を基に検出統計量を求め外れ値を求める。

【0048】第16の発明においては、入力された値が複数のグループに分類されていて、複数の要因により判別分析を行う場合に、判別関数値を計算し計算された値を基に検出統計量を求め外れ値を求める。

【0049】第17の発明におけるデータ処理装置は、計測手段によりN個の値を計測し、この計測された値から、上記外れ値検出方法を実行する外れ値検出手段により、外れ値を検出し、出力手段により外れ値を知らせる。

【0050】第18の発明におけるデータ処理装置は、計測手段によりN個の値を計測し、この計測された値から、上記外れ値検出方法を実行する外れ値検出手段により、外れ値を検出し、データ処理手段により検出された外れ値を除いた残りの値を用いて所定の処理を実行する。

【0051】

【実施例】

実施例1. 従来例で説明した図13を再びこの実施例の

装置を説明するための図として説明する。図13で、1は情報処理装置、2はコンピュータ(FDD付)、3はディスプレイ・ユニット、4はプリンタ、5はキーボード、6はフロッピーディスクである。この発明のハードウェア構成は従来例と変わらず、プログラム・ルーチンが記憶されたフロッピーディスク6をコンピュータ(FDD付)2に挿入し、オペレーション・ソフトを駆動して、情報処理装置1をスタートさせる。プログラム・ル*

$$U_t = n \log \sigma + 2S$$

ただし、 n はサンプル数

S は外れ値の候補の個数

σ は x_1, x_2, \dots, x_n をサンプルデータとすると、

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

【0054】この統計量の値が最小になるサンプルの組み合わせを見つければよい。図1は本発明の説明のためのフローチャートである。ステップ20は、キーボード5からのデータを連続的に入力する入力工程である。例えば、 x_1, x_2, x_3, x_4, x_5 の5つのデータを入力する。この場合は、入力するデータの個数を N とすると、 $N=5$ となる。ステップ21は、入力されたデータの大きさを比較し、例えば昇順に $x_1 < x_2 < x_3 < x_4 < x_5$ のように並べる。この工程は大小判定工程である。

【0055】このようにデータを昇順に並べかえることによって、外れ値の候補を見つけることが容易となる。外れ値の候補の個数を s ($s \geq 1$) とすると、外れ値の候補は、その性質からいって一番大きい値から s 個、一番小さい値から s 個、または大きい値と小さい値の両方あわせて s 個と考えられる。

【0056】ステップ22はこれらのデータ群から、本実施例での計算式により検出統計量 U_t を計算する算出工程である。外れ値の候補の個数 $s=1$ の場合は、まず、(x_1, x_2, x_3, x_4, x_5)から x_1 を除いた時の検出統計量を計算する。これを検出統計量 $U_{t(-1)}$ とする。以下同様に x_5 を除いた時を $U_{t(-5)}$ とする。外れ値の候補の個数 $s=2$ の場合は、 x_1 と x_2 を除いた時を $U_{t(-1,-2)}$ とし、 x_4 と x_5 を除いた時を $U_{t(-4,-5)}$ とし、 x_1 と x_5 を除いた時を $U_{t(-1,-5)}$ とし、小さい値または大きい値から順にサンプルを除いて検出統計量を計算する。このように考えられる組み合わせの検出統計量をそれぞれ計算する。ここで外れ値の候補の個数 s は、予めシステムにより定められているものとする。あるいは、外れ値の候補の個数 s は、オペレータ、あるいは、プログラムにより指定されるものとする。あるいは、外れ値の候補の個数は計算の度に自

*ーチンがロードされ、入力待状態となる。キーボード5からデータをキー入力すれば、プログラム・ルーチンが動作し、ディスプレイ3に処理結果を表示し、また、プリンタ4に処理結果をプリントすることになる。

【0052】この実施例では、検出統計量を算出するために $\log 2$ を使う。

【0053】

【 $\log 2$ 】

由に設定することが可能なものであるとする。

【0057】尚、ここで与えられる外れ値の候補の個数は、外れ値として必ず見つけなければならない個数ではない。ここで言う外れ値の候補の個数とは、外れ値として検出する最大の個数を言う。例えば、外れ値の候補の個数 $s=2$ の場合は、外れ値を最大2個見つける場合を言い、外れ値の個数を必ず2個見つけるという意味ではない。従って、外れ値の候補の個数 $s=2$ の場合は、外れ値が0個の場合、外れ値が1個の場合、あるいは、外れ値2個の場合というような結果が考えられる。以下同様に外れ値の候補の個数 s という場合は、外れ値として検出できる数の最大値を示すものとする。このように、この実施例及び後述する実施例においては、外れ値の数を特定の数に設定する必要はなく、外れ値の数の最大値を指定しておけばよい。

【0058】ステップ23は、検出統計量 U_t の最小値(U_{tmin})を見つける段階である。ステップ24は、外れ値の候補を除かない時の検出統計量 $U_{t(0)}$ と U_{tmin} を比較する段階である。ステップ25は、 U_{tmin} の方が小さい場合、 U_{tmin} を求めた時のデータの組み合わせに含まれていなかった値を外れ値とする。ステップ26は、 $U_{t(0)}$ が最小となる場合で、この時、外れ値は「ない」とする。ステップ23からステップ26までが外れ値検出工程である。ステップ27は、表示等をする。

【0059】次に、データを使ってこのフローチャートの流れを説明する。ステップ20で次の5つのデータを入力する。

5. 71, 6. 57, 7. 29, 8. 06, 13. 32
ステップ21で入力されたデータを昇順に並べかえる。ステップ22で $\log 2$ を使い、 U_t の値を計算する。例えば、一番小さな値5. 71を除いた時は、サンプル数 n

=4、外れ値の候補の個数 $s=1$ であるので、検出統計量を $U_{t(-1)}$ と表すと、 $U_{t(-1)}=5.908$ となる。一番大きな値 13.32 を除いた時の検出統計量は、サンプル数 $n=4$ 、外れ値の候補の個数 $s=1$ であるので、 $U_{t(-5)}$ と表すと、 $U_{t(-5)}=1.440$ となる。5.71 と 13.32 をともに除いた時は、サンプル数 $n=3$ 、外れ値の候補の個数 $s=2$ であるので、 $U_t *$

サンプルの組み合わせと検出統計量の値

外れ値 の候補 大きい 小さい	なし	13.32	13.32 8.06
なし	4.930 ($S=0$)	($S=1$) 1.440 (は小値)	2.689 ($S=2$)
5.71	5.71 ($S=1$)	2.509 ($S=2$)	3.957
5.71 6.57	6.957 ($S=2$)	4.091	

【0061】ステップ23で以上で求めた U_t の最小値 U_{tmin} を求めると、 $U_{tmin}=U_{t(-5)}=1.440$ である。ステップ24で外れ値の候補を除かない時の検出統計量 $U_{t(0)}=4.930$ を計算し、 U_{tmin} と $U_{t(0)}$ を比較する。すると、 $U_{t(-5)} > U_{t(0)}$ は成立しないので、ステップ25へ行き、 $U_{t(-5)}$ の時のデータの組み合わせを外れ値とする。即ち、13.32 が外れ値とわかる。ステップ27で、外れ値 13.32 の表示等出力を行う。

【0062】このように外れ値の候補の個数が3の場合であっても、検出された外れ値の個数は1つであり、外れ値の候補の個数以内の範囲で外れ値を検出することが

【0063】この例で示した 5.71、6.57、7.29、8.06、13.32 のデータの場合、13.32 を外れ値とするのは、竹内 啓 (1980) 「現象と行動の中の統計数理」新曜社でも同様の結果となっている。

【0064】この実施例の図2に示した検出統計量の式 $U_t = n \log \sigma + 2s$ は、AIC (AKAIKE'S information criterion) のアナロジーから考えられ

* $(-1, -5)$ と表すと、 $U_{t(-1, -5)}=2.509$ となる。また、 $U_{t(-1, -2)}$ 、 $U_{t(-4, -5)}$ 、 $U_{t(-1, -2, -5)}$ 、 $U_{t(-1, -4, -5)}$ を計算すると (即ち、外れ値の候補の個数 $s=3$ の検出統計量 $U_{t(0)}$ を計算すると) 表2のようになる。

【0060】

【表2】

た。 n はサンプル数、 s は外れ値の候補の個数、 σ^2 は分散、 σ は標準偏差である。この式の第1項は、外れ値の候補であるサンプルが除かれると小さくなる傾向がある。というのは、分散 σ^2 は外れ値を除くと小さくなるからである。また、サンプル数 n も外れ値の候補として除く数が増えると、 $n = (\text{データ数 } N) - (\text{外れ値の候補の個数 } s)$ であるから、例えば、 n は5から4、5から3というように小さな値になるからである。第2項は、サンプル数が多くなると増加する。従って、外れ値を除いた時、第1項と第2項の和 U_t は最小になると考えることができる。

【0065】次に、この関係を Grubbs のデータ1を使って述べる。Grubbs のデータ1を次に示す。2.02、2.22、3.04、3.23、3.59、3.73、3.94、4.05、4.11、4.13
総データ数は10個である。このデータから $\log \sigma$ を計算した値を表3に、 $n \log \sigma$ を計算した値を表4に示し、 $U_t = n \log \sigma + 2s$ を計算した値を表5に示す。

【0066】

【表3】

Grubbsのデータ1による $\log \sigma$ の値

外れ値の候補 大 小	なし	4.13	4.13 4.11	4.13 4.11 4.05
なし	-0.313	-0.318	-0.337	-0.376
2.02	-0.514	-0.516	-0.533	—
2.02 2.22	-0.95	-0.967	—	—
2.02 2.22 3.04	-1.183	—	—	—

【0067】

* * 【表4】

Grubbsのデータ1による $n \log \sigma$ の値

外れ値の候補 大 小	なし	4.13	4.13 4.11	4.13 4.11 4.05
なし	-3.127	-2.858	-2.697	-2.629
2.02	-4.630	-4.127	-3.732	—
2.02 2.22	-7.6	-6.771	—	—
2.02 2.22 3.04	-8.284	—	—	—

【0068】

* * 【表5】

Grubbsのデータ1による $U_t = \log \sigma + 2s$ の値

外れ値の候補 大きい 小さい	なし	4.13	4.13 4.11	4.13 4.11 4.05
なし	(S=0) -3.127	(S=1) -0.858	(S=2) 1.303	(S=3) 3.371
2.02	(S=1) -2.630	(S=2) -0.127	(S=3) 2.268	—
2.02 2.22	(S=2) -3.600 最小	(S=3) -0.771	—	—
2.02 2.22 3.04	(S=3) -2.284	—	—	—

【0069】 これらをグラフにしたものが図2である。

図2のx軸は、外れ値の候補の個数 s であり、y軸は U_t の値である。 s に対応する U_t の値が複数ある場合は、その中の最小のものをプロットした。例えば、表3において外れ値の候補の個数 $s=1$ の場合、 U_t は-0.318と-0.514となるが、-0.514を用いてプロットした。図2の一点鎖線で示したグラフ(1)は、 $U_t = \log \sigma$ とした場合を示している。点線で示したグラフ(2)は、 $U_t = n \log \sigma$ とした場合を示している。実線で示したグラフ(3)は、 $U_t = 2s$ とした場合である。太線で示したグラフ(4)は、 $n \log \sigma$ と $2s$ を加算した $U_t = n \log \sigma + 2s$ の

値である。

【0070】 前述したように外れ値の候補の個数 s が増加するに従って、第1項の $n \log \sigma$ は減少することがグラフ(2)よりわかる。そして、第2項の $2s$ は増加することが、グラフ(3)よりわかる。グラフ(4)に示す $n \log \sigma + 2s$ の値は、 $s=2$ で最小値を取った後増加している。 $U_t = n \log \sigma + 2s$ の場合、最小値は1ヶである。表5からわかるように、グラフ(4)が最小となるのは、外れ値の候補の個数 $s=2$ であって、その外れ値として2.02と2.22を仮定した場合である。これは、Kitagawaの方法でも同じ結果を得ている(Genshiro Kitagawa (1

979): "On the Use of Aic f or the Detection of Outliers", Technometrics, Vol. 2 1, No. 2)。

【0071】次に、 $U_t = n \log \sigma + 2s$ と s の関係をもう1つ別の例で述べる。データとして、
-1.40, -0.44, -0.30, -0.24, -*

*0.22, -0.15, -0.13, 0.06, 0.1
0.018, 0.20, 0.48, 0.63, 1.0
1

とする。これにより得られた $U_t = n \log \sigma + 2s$ の値を表6に示す。

【0072】

【表6】

外れ値の候補 の候補 大きい 小さい	なし	1.01	1.01 0.63	1.01 0.63 0.48
なし	-9.42	-8.32	-6.24	-3.89
-1.40	-11.15	-11.12	-7.65	—
-1.40 -0.44	-8.82	-8.81	—	—
-1.40 -0.44 -0.30	-6.14	—	—	—

【0073】表6を基に、グラフを書くと同図3のようになる。図3のx軸は外れ値の候補の個数 s 、y軸は $U_t = n \log \sigma + 2s$ の値である。図3からわかるように、この場合も U_t の値が最小になった後、 s が大きくなるにつれ、 U_t の値も大きくなっている。また、外れ値の性格より、外れ値の数は総データ数に比して小さな数であると考えられる。よって、以後検出統計量の計算結果を表に示す場合、最小の前後のデータのみを示すことにする。この外れ値の性格を用いることにより、外れ値の候補の数を予め指定することなく、外れ値を検出することも可能である。前述したように、外れ値の数が大きくなるにつれて検出統計量の値も大きくなる。従って、外れ値の候補の数を指定しない場合には、外れ値の候補の数が少ない順に検出統計量を算出し、順に外れ値の候補の数を増やして検出統計量を算出し、その計算した検出統計量が次第に大きくなる場合には、その計算を終了させる。このようにして、外れ値の候補の数が予め指定されていない場合であっても、外れ値を検出することが可能になる。従って、前述したように外れ値の候補の数を予め指定する場合以外に、外れ値の候補の数をシステムやプログラムにより指定せずに、検出統計量の計算結果を比較していくことにより、その計算結果が次第に大きくなることが判明した時点で検出統計量の計算を停止させることにより、外れ値を検出することが可能になる。次に、検出統計量 U_t が有効であるかどうか検証するために、従来の計算方法による結果と比較したものを実施例2から実施例6で述べる。

【0074】実施例2. この実施例では、Grubbsのデータ2を用い、検出統計量の式として U_t を用いた場合の外れ値について述べる。Grubbsのデータは、全て次の文献より引用している。

"Procedures for detecting

outlying Observations in samples", Technometrics, Vol. 11, 1-21

Grubbsのデータ2は次の値である（データ数は12）。

0.745, 1.832, 1.856, 1.884,
1.914, 1.916, 1.947, 1.949,
2.013, 2.023, 2.045, 2.327

原典では、3回の観測値とその平均値が載っているが、ここでは平均値のみを昇順に載せる。検出統計量の計算結果を表7に示す。

【0075】

【表7】

Grubbsのデータ2と検出統計量

検出統計量	$U_t = n \log \sigma + 2s$
0.745	-2.214
2.327	-9.578
2.045 2.327	-8.244
0.745	-20.497(2)
0.745 1.832	-18.610
0.745 2.327	-22.861(1)
0.745 1.832 2.327	-19.107
0.745 2.045 2.327	-19.145

【0076】表7より U_t が最小値をとるのは、0.745と2.327を外れ値とした場合である。前述したKitagawaの方法も同じ結果となっている。

【0077】実施例3. この実施例は、Grubbsのデータ3を用いた場合について述べる。Grubbsのデータ3は次の値である（データ数は10）。

568, 570, 570, 570, 572, 572, 572, 578, 584, 596

検出統計量の計算結果を表8に示す。

【0078】

【表8】

Grubbsのデータ3 (総データ数10) と検出統計量

検出統計量	$n \log s + 2s$
596	21.109
584 596	15.975(2)
568	12.191(1)
568 570	21.075
568 596	21.158
	16.319

【0079】表8より U_t が最小値をとるのは、584、596を外れ値とした場合である。Grubbsによると596を外れ値としている。Dallas E.

Johnson他は、584、596を外れ値とした。これは、数2で求めた場合と同じである。なお、以後Dallas E. Johnson他という場合は、次の資料に基づくものとする。

Dallas E. Johnson, Stephen A. McGuire, and George A. Milliken (1978): "Estimating σ^2 in the Presence of Outliers", Technometrics, Vol. 20, No. 4

【0080】実施例4. 更に、実施例3のデータで、同じデータを重複させてサイズを2倍にしたものを用いた場合を次に示す。データ4は次の値である (データ数は20)。

568, 568, 570, 570, 570, 570, 570, 570, 572, 572, 572, 572, 572, 572, 578, 578, 584, 584, 596, 596

【0081】この実施例では、データのサイズを2倍にしたので、外れ値の候補の個数を4 ($s=4$) とする場合について説明する。外れ値の候補の個数が4の場合は、以下のような組み合わせに対して検出統計量を算出することになる。即ち、外れ値の候補の個数が1 ($s=1$) の場合の検出統計量と、外れ値の候補の個数が2 ($s=2$) の場合の検出統計量と、外れ値の候補の個数が3 ($s=3$) の場合の検出統計量と、外れ値の候補の個数が4 ($s=4$) の場合の検出統計量を求める必要がある。外れ値の候補の個数 s に対応する検出統計量は、以下に示すとおりである。

$s=1 \quad U_t(-1)$

$U_t(-20)$

$s=2 \quad U_t(-1, -2)$

$U_t(-1, -20)$

$U_t(-19, -20)$

$s=3 \quad U_t(-1, -2, -3)$

$U_t(-1, -2, -20)$

$U_t(-1, -19, -20)$

$U_t(-18, -19, -20)$

$s=4 \quad U_t(-1, -2, -3, -4)$

$U_t(-1, -2, -3, -20)$

$U_t(-1, -2, -19, -20)$

$U_t(-1, -18, -19, -20)$

10 $U_t(-17, -18, -19, -20)$

【0082】外れ値の候補の個数が4の場合においても図1に示したフローチャート同様の順に外れ値を検出することが可能である。異なる点は、図1におけるステップ22において前述したような $s=1$ から $s=4$ までのそれぞれの検出統計量を算出する点である。このようにして、算出された検出統計量 U_t の計算結果を表9に示す。

【0083】

【表9】

20 Grubbsのデータ4 (総データ数20) と検出統計量

検出統計量	$n \log s + 2s$
596	42.218
596 596	38.732
584 596 596	31.950
584 584 596 596	29.538(2)
568	24.382(1)
568 568	42.207
568 568 570	42.150
568 568 570 570	42.253
568 596	42.317
568 596 596	38.887
568 596 596 596	32.330
568 568 596 596	30.055
568 568 596	38.993
568 568 570 596	39.294

【0084】表9より外れ値は、584、584、596、596である。Dallas E. Johnson他も同様の結論となっている。

40 【0085】尚、外れ値の候補の個数は、入力されたデータの数に基づいて常識的な範囲で任意に設定できるものである。例えば、入力されたデータの数が5 ($N=5$) である場合に、外れ値の候補の個数は1又は2 ($s=1$ 又は2) とするのが常識的な範囲である。また、入力されたデータの数が多くなれば外れ値の候補の個数も多くする分には差し支えない。このように外れ値の候補の個数は、入力されたデータの数、あるいは、そのシステムにおいて、どの位の精度を要求しているかというシステムの要求に応じて判断されるべきものである。前述

50 した実施例、あるいは、後述する実施例においては、外

れ値の数を何個と推定するかという判断は、予めシステムにより定められているか、あるいは、オペレータやプログラムにより任意に指定できるものとする。

【0086】実施例5. 次に、Rosnerのデータを用いた例について述べる。この例は、サイズが54と比較的大きく、外れ値も多く存在すると考えられるケースである。次に、データを示す。

-0.25, 0.68, 0.94, 1.15, 1.20, 1.26, 1.26, 1.34, 1.38, 1.43, 1.49, 1.49, 1.55, 1.56, 1.58, 1.65, 1.69, 1.70, 1.76, 1.77, 1.81, 1.91, 1.94, 1.96, 1.99, 2.06, 2.09, 2.10, 2.14, 2.15, 2.23, 2.24, 2.26, 2.35, 2.3*

Rosnerのデータと検出統計量

外れ値として検出したサンプル	検出統計量
なし	8.564
-0.25	8.379
-0.25 0.68	9.607
6.01	5.415
5.42 6.01	3.008
5.34 5.42 6.01	-0.234
-0.25 6.01	4.939
-0.25 5.42 6.01	2.225
-0.25 0.68 5.42 6.01	3.346
4.64 5.34 5.42 6.01	-2.009
4.30 4.64 5.34 5.42 6.01	-3.305(1)
3.68 4.30 4.64 5.34 5.42 6.01	-2.995(2)
3.59 3.68 4.30 4.64 5.34 5.42 6.01	-2.681
3.30 3.59 3.68 4.30 4.64 5.34 5.42 6.01	-1.720
-0.25 5.34 5.42 6.01	-1.518

【0088】Rosnerは、外れ値が最大10個と仮定して検定した。危険率5%で、5.34、5.42、6.01を外れ値とした。表10は、この実施例による検出統計量の計算結果を示す表である。前述したように検出統計量の計算結果を表に示す場合には、最小の値の前後のデータのみを示してある。この表10からわかるように、計算式 U_t を用いた場合、外れ値は4.30、4.65、5.34、5.42、6.01である。この場合には、外れ値として5つの外れ値が検出されているが、Rosnerが仮定したように外れ値が最大10個あると仮定した場合であっても、あるいは、外れ値の候補の数を指定せずに外れ値の候補の数を増やす毎に計算された検出統計量と比較することにより自動的に外れ値を検出した場合のいずれの場合においても、結果は

*7.240, 2.47, 2.54, 2.62, 2.64, 2.90, 2.92, 2.92, 2.93, 3.21, 3.26, 3.30, 3.59, 3.68, 4.30, 4.64, 5.34, 5.42, 6.01

この、Rosnerのデータは次の文献からとった。

Bernard Rosner (1977): "Percentage Point for a Generalized ESD Many-Outlier Procedure", Technometrics, Vol. 25, No. 2

次に、 U_t の計算結果を表10に示す。

【0087】

【表10】

この5つの外れ値を検出する。この実施例においては、5つの外れ値を検出したが、もしこの方法で外れ値が3個までとすると、Rosnerと一致している。

【0089】実施例6. 正規乱数、指数乱数、一様乱数をサンプルデータとした場合について述べる。正規乱数、指数乱数は、外れ値が現れる可能性があるが、一様乱数からは外れ値は出て欲しくない。正規乱数データは ($n=10, \sim N(0, 1)$) より、

-2.666, -1.272, -0.042, 0.140, 0.273, 0.415, 0.467, 1.160, 1.672, 1.673

である。 U_t の計算結果を表11に示す。

【0090】

【表11】

正規乱数と検出統計量データ(n=10, ~N(0,1))

外れ値として除いたサンプル	検出統計量	nlogσ+2s
なし		2.294
1.167		3.475
1.672 1.673		4.998
-2.666		0.744(2)
-2.666 -1.272		0.435(1)
-2.666 1.673		2.320
1.160 1.672 1.673		6.546
-2.666 1.673		3.329
-2.666 -1.272 1.673		2.017
-2.666 -1.272 -0.042		2.559

【0091】表11より外れ値は-2.666、-1.272である。

【0092】次に、指数乱数を用いた場合について示す。データは、竹内「現象と行動の中の統計数理」(新曜社)からとった。

0.003, 0.021, 0.161, 0.178,
0.180, 0.210, 0.249, 0.413,
0.494, 0.562, 0.613, 0.879,
0.981, 1.059, 1.131, 1.264,
2.367, 3.669, 3.826, 4.193

総データ数は20である。Utの計算結果を表12に示す。

【0093】

【表12】

指数乱数と検出統計量

外れ値として除いたサンプル	検出統計量	nlogσ+2s
なし		5.041
4.193		3.894
4.193 3.826		2.074
4.193 3.826 3.669		-3.091(2)
4.193 3.826 3.669 2.367		-6.402(1)
0.003		6.889
0.003 0.021		8.706
0.003 0.021 0.161		10.573
0.003 0.021 0.161 0.178		12.405
4.193 0.003		5.896
0.003 4.193 3.826		4.261
0.003 4.193 3.826 3.669		-0.674
0.003 0.021 4.193		7.886
0.003 0.021 4.193 3.826		6.415

【0094】表12より外れ値は、4.193、3.826、3.669、2.367である。

【0095】次に、[0, 1]の一樣乱数を用いた場合について述べる。データ数10でデータは次の通りである。

0.283, 0.470, 0.643, 0.688,
0.916, 0.930, 0.945, 0.953,
0.973, 0.995

Utの計算結果を表13に示す。

10 【0096】

【表13】

[0, 1]の一樣乱数と検出統計量

外れ値として除いたサンプル	検出統計量	nlogσ+2s
なし		-14.484(1)
0.995		-11.003
0.995 0.973		-7.536
0.283		-13.649(2)
0.283 0.470		-12.533
0.283 0.995		-9.878
0.283 0.470		-4.109
0.283 0.995 0.973		-6.103
0.283 0.470 0.995		-8.435
0.283 0.470 0.643		-10.439

【0097】表13より外れ値の候補がない場合が最小となっているので、一樣乱数の場合、外れ値の候補はない。一樣乱数については、更に1ケース試みたが同様に外れ値の候補はなかった。一樣乱数という性格上、外れ値の候補なしということは望ましい結果である。

30 【0098】実施例7。従来の技術に出ているデータについて、Utを用いて検出統計量を求める。データは、5.71, 6.57, 7.29, 8.06, 10.00, 15.00

である。結果は表14のようになる。

【0099】

【表14】

サンプルの組み合わせと検出統計量の値

ケース	n	s	検出統計量
除かないとき	6	0	6.77 $U t(0)$
5.71を除いたとき	5	1	7.55 $U t(-1)$
5.71と15.00 ともに除いたとき	4	2	4.99 $U t(-1,-6)$
15.00を除いたとき	5	1	3.90 $U t(-6)(84)$

【0100】表14より $U t$ は15.00を除いた時、最小値となることがわかり、15.00を外れ値とする。

【0101】実施例8. 従来技術で述べたマスク効果のデータについて、数2に示した計算式を用いて検出統計量を求める。データは、

* 5.71, 6.57, 7.29, 8.06, 14.80, 15.00

である。結果は表15のようになる。

【0102】

【表15】

*

サンプルの組み合わせと検出統計量の値

ケース	n	s	検出統計量
除かないとき	6	0	8.06 $U t(0)$
5.71を除いたとき	5	1	8.61 $U t(-1)$
5.71, 15.00 を除いたとき	4	2	8.76 $U t(-1,-6)$
15.00を除いたとき	5	1	7.90 $U t(-6)$
15.00, 14.80 を除いたとき	4	2	3.44 $U t(-6,-5)$
5.71, 6.57 を除いたとき	4	2	9.15 $U t(-1,-2)$
5.71, 15.00, 14.80を除いたとき	3	3	4.51 $U t(-1,-6,-5)$
5.71, 6.57, 15.00を除いたとき	3	3	9.65 $U t(-1,-2,-6)$

【0103】表15より $U t(-6,-5)$ の時、最小値となることがわかり、15.00, 14.80を外れ値とする。このように数2に示した計算式を用いれば、マスク効果を回避することができる。

【0104】実施例9. 図4は、この実施例を説明するための図である。この実施例のデータ処理装置は、センサー等の計測手段を有し、これより得られた測定値から、外れ値を検出する外れ値検出手段を有す。次に、外

れ値がある場合は、これを除いたデータで平均値を求めるデータ処理手段を有す。次に、実際の適用例について述べる。センサーから得られる測定値を、一定時間間隔ごとに5個測定し外れ値を検出し、これを除いた平均値を測定値とすることを考える。 x_1, x_2, x_3, x_4, x_5 が、測定データとして得られる。検出統計量 $U t$ を計算し、外れ値を求める。外れ値があれば外れ値を除き偏りのない平均値を求めることができる。データ

は時刻 t_1 、 t_2 、 t_3 、 t_4 について表16に示す。

※【表16】

【0105】

時刻	データ					平均値
t_1	3.23	1.24	2.03	2.86	1.02	2.08
t_2	2.06	0.74	2.19	2.39	2.73	2.01
t_3	2.90	0.08	2.55	2.23	2.53	2.06
t_4	1.83	2.36	3.36	2.44	1.95	2.39

【0106】時刻 t_1 のデータについて検出統計量 U_t を求める。除かない時、 $U_t(0) = -0.711$ となる。1.02を除いた時、 $U_t(-1) = 0.952$ 、3.23を除いた時、 $U_t(-5) = 0.709$ となる。1.02、3.23ともに除いた時、 $U_t(-1, -5) = 2.760$ となる。これを表17にまとめると次のようになる。 $U_t(-1, -2)$ 、 $U_t(-4, -5)$ も考えられるがこの実施例では影響がないので表に示すのを省略する。以下の表でも同様に影響がないものは表示しないことにする。

【0107】

【表17】

外れ値の 候補	大きい方	なし	3.23
	小さい方		
	なし	-0.711	0.709
	1.02	0.952	2.760

【0108】外れ値がない時の U_t が -0.711 と最小である。従って外れ値はない。時刻 t_2 のデータについて U_t を求め、表18に示す。

【0109】

【表18】

外れ値の 候補	大きい方		
	小さい方	なし	2.73
	なし	-1.969	0.205
	0.74	-3.524	-2.504

時刻	更新されたデータ					更新された 平均値
t_1	3.23	1.24	2.03	2.86	1.02	2.08
t_2	2.06	2.19	2.36	2.73		2.33
t_3	2.90	2.55	2.23	2.53		2.55
t_4	1.83	2.36	2.44	1.95		2.15

【0116】実施例10. この実施例は、外れ値検出手段を有し、外れ値を知らせる出力手段を有するデータ処理装置について述べる。入力工程から表22のようなデータが得られ、このデータから U_t を計算することによ

※【0110】0.74を除いた時の U_t が -3.524 と最小である。従って0.74を外れ値とする。時刻 t_3 のデータについて U_t を求め、表19に示す。

【0111】

【表19】

外れ値の 候補	大きい方		
	小さい方	なし	2.90
	なし	0.057	2.112
	0.08	-3.753	-1.765

【0112】従って、0.08を外れ値とする。次項 t_4 のデータについて U_t を求め、表20に示す。

【0113】

【表20】

外れ値の 候補	大きい方		
	小さい方	なし	3.36
	なし	-3.092	-3.888
	1.36	-0.652	-0.616

【0114】従って、3.36を外れ値とする。更新されたデータ及び平均値は、表21のようになる。

【0115】

【表21】

り外れ値を検出する。表22のデータを図示すると図5のようになる。

【0117】

【表22】

	27										28			
t	1	2	3	4	5	6	7	8	9	10	11	12	13	14
y	1	1.5	-1	-2	-1.5	0	-3	0	1.5	1	4	0	-1	0

【0118】このデータについて検出統計量 U_t を求め * 【0119】
ると表23のようになる。 * 【表23】

外れ値の 検出	大きい方 小さい方	なし	4	4, 1.5
なし		7.30	5.65	6.89
-3		7.39	5.23	6.54
-3, -2		8.22	5.77	6.54

【0120】表23より4、-3を外れ値とする。従来は、入力工程から得られたデータをディスプレイ等に図5に示すようなグラフを表示し、人が目視によって外れ値と思われる値をピックアップしてから計算し、確かめていた。本発明による装置により外れ値を自動的に検出し、工程環境・条件に異常があったかどうかの確認を行うことができる。

【0121】実施例11. 図6はこの実施例を説明するための図である。図6で示した外れ値検出方法は、図1の入力工程と大小判定工程の間に、入力したデータを加工する加工工程が追加されたものである。この実施例では、時間とともに増加する傾向がある特性値を出力する装置からデータをうけとり、加工工程により時間とともに増加する傾向を除いたデータから、外れ値を検出する装置について述べる。データを表24に示す。これを図示すると図7の様になり、データは時間とともに増加する傾向があることが分かる。

【0122】

【表24】

t	1	2	3	4	5	6	7	8	9	10
y	2	4	3	6	8	5	6	11	6	9

【0123】そこで、最小2乗法により傾向直線を求めると次のようになる。

$$y = 1.93 + 0.70t$$

データから傾向直線の値を引くことにより、下のようなデータとなる。

-0.63, 0.67, -1.03, 1.27, 0.57, -1.13, -0.83, 3.47, -2.23, 0.07

これを図示すると、図8のようになる。この補正されたデータから、 U_t を計算すると表25のようになる。

【0124】

【表25】

外れ値の 検出	大きい方 小さい方	なし	3.47
なし		4.13	2.29
-2.23		4.92	2.83

【0125】従って、補正されたデータ3.47に対応する $t=8$ の $y=11$ を外れ値とする。この様に、時間とともに増加する傾向のあるデータから、増加する傾向を補正することにより、実施例1の外れ値検出方法を適用することができる。

【0126】実施例12. この実施例は、一つのサンプルに複数の特性値がある場合に、加工工程においてデコ比 $(X(X^T X)^{-1} X^T)$ の対角要素を計算し、これをもとに外れ値を検出する装置について説明する。データおよび計算されたデコ比を表26に示す。

【0127】

【表26】

29

no.	x1	x2	x3	x4	テコ比
1	7	26	6	60	0.48
2	1	29	15	52	0.28
3	11	58	8	20	0.13
4	11	31	8	47	0.24
5	7	52	6	33	0.34
6	11	55	9	22	0.12
7	3	71	17	6	0.36
8	1	31	22	44	0.38
9	2	54	18	22	0.19
10	21	47	4	26	0.67
11	1	40	23	34	0.37
12	11	68	9	12	0.19
13	10	68	8	12	0.24
14	30	85	30	79	0.99

30

*【0128】テコ比を用いて検出統計量Utを求めると表27のようになる。

【0129】

【表27】

10

*

外れ値の候補 大きい方 小さい方	なし	0.99	0.99
なし	-20.75	-22.99	-22.85
0.12	-17.37	-19.50	-19.39
0.12, 0.13	-14.02	-16.08	-16.12

【0130】-22.99が最小値である。従って0.99つまりサンプルno. 14を外れ値とする。テコ比が大きいデータは、全体に与える影響が大きいので、外れ値か否か容易に判定できる外れ値検出装置があることは有効である。

【0131】実施例13. この実施例は、入力されたデ※

※一タが回帰分析のモデルの場合、加工工程において回帰分析の残差を求め、これをもとに外れ値を検出する装置について説明する。データおよび回帰式により求めた残差は表28のようになる。

【0132】

【表28】

no.	x1	x2	x3	y	残差
1	7	26	50	78.5	0.346
2	1	29	52	74.3	1.545
3	11	56	20	104.3	-1.874
4	11	31	47	87.6	-1.783
5	7	52	33	95.9	-0.322
6	11	55	22	109.2	3.948
7	3	71	6	102.7	-1.339
8	1	31	44	72.5	-3.186
9	2	54	22	93.1	1.288
10	21	47	26	115.9	0.246
11	1	40	34	83.8	1.993
12	11	68	12	113.3	1.171
13	10	68	12	109.4	-2.033

【0133】残差を用いて検出統計量Utを求めると表29のようになる。

★【0134】

★【表29】

外れ値の候補 大きい方 小さい方	なし	3.948	3.948
なし	8.459	7.723	8.573
-3.186	8.762	7.873	8.816
-3.186, -2.033	9.746	8.939	10.030

【0135】Ut 7.723が最小である。従って残差 50 3.948つまりサンプルno. 6が外れ値となる。

【0136】実施例14. この実施例は、入力されたデータの特性値が複数あり正準相関分析モデルを適用できる場合、加工工程において次に示すようにデータを加工し、これを用いて外れ値を検出する装置について説明する。表30のようなデータについて考える。

【0137】

【表30】

no.	特性値		要因			
	y1	y2	x1	x2	x3	x4
1	78.5	21.0	7	26	6	60
2	74.3	22.0	1	29	15	52
3	104.3	23.0	11	56	8	20
4	87.6	24.0	11	31	8	47
5	95.9	23.1	7	52	6	33
6	109.2	19.6	11	55	9	22
7	102.7	18.7	3	71	17	6
8	72.5	23.3	1	31	22	44
9	93.1	23.8	2	54	18	22
10	115.9	19.4	21	47	4	26
11	83.8	28.8	1	40	23	34
12	113.3	19.8	11	66	9	12
13	109.4	19.1	10	68	8	12

【0138】このデータに正準相関分析を行い、y1, y2の合成変量関数が2個求まる。この合成変量関数を用いて合成変量関数値が求まる。合成変量関数値をもとにテコ比を計算する。テコ比は次のようになる。

0.27, 0.31, 0.18, 0.12, 0.10, 0.20, 0.19, 0.30, 0.12, 0.30, 0.63, 0.13, 0.16

このテコ比について検出統計量U_tを計算すると表31のようになる。

【0139】

【表31】

外れ値 小さい方	大きい方	
	なし	0.63
なし	-26.00	-29.03
0.10	-22.03	-24.98

【0140】従ってテコ比0.63を外れ値とする。これはno. 11のサンプルである。

【0141】実施例15. この実施例は、入力されたデータが2グループに特性値が分類されていて、加工工程において複数の要因により判別分析を行い、この加工されたデータをもとに各グループでの外れ値を検出する装置について述べる。データは表32に示す。

【0142】

【表32】

データ		x1	x2	判別関数値
no.				
グループ1	1	6	0	-3.80
	2	0	2	-0.56
	3	0	3	-0.84
	4	1	2	-1.20
	5	1	3	-1.48
	6	1	4	-1.76
グループ2	7	4	0	-2.53
	8	4	1	-2.82
	9	5	0	-3.17
	10	5	1	-3.45
	11	5	2	-3.73
	12	0	4	-1.12

【0143】データをプロットすると図9のようになる。図よりグループ1では、サンプルno. 1が、グループ2ではサンプルno. 12が外れ値のようである。データについて判別分析を実施し、判別関数値を求めると、

$$y = -0.634 * x1 - 0.281 * x2$$

となる。この判別関数を用いて判別関数値を計算すると例えば、no. 1の場合、

$$\text{判別関数値} = -0.634 * 6 - 0.281 * 0 = -3.80$$

となる。判別関数値を表32の右欄に載せた。グループ1の判別関数値についてU_tを求めると表33のようになる。

【0144】

【表33】

外れ値 小さい方	大きい方	
	なし	-0.56
なし	0.33	2.19
-3.80	-2.24	-0.33

【0145】従って判別関数値-3.80、サンプルno. 1を外れ値とする。グループ2の判別関数値についてU_tを求めると表34のようになる。

【0146】

【表34】

外れ値 小さい方	大きい方	
	なし	-1.124
なし	-1.00	-2.24
-3.73	0.94	-0.24

【0147】従って判別関数値-1.12、サンプルno. 12を外れ値とする。

【0148】実施例16. この実施例は、入力されたデータが3グループに特性値が分類されていて、加工工程において複数の要因により判別分析を行い、このデータ

にもとづいて各グループでの外れ値を検出する装置について述べる。データを表35に示す。

データ			* 【0149】 【表35】		
no.	x1	x2	判別関数値 f_1	判別関数値 f_2	ユークリッド距離 d
グループ1	1	0	3.45	-1.21	3.65
	2	0	4.60	-1.62	4.88
	3	1	1.46	-1.33	1.97
	4	1	2.61	-1.74	3.14
	5	1	3.76	-2.14	4.33
グループ2	6	4	-3.39	-2.11	3.99
	7	5	-4.23	-2.63	4.98
	8	4	-2.24	-2.51	3.36
	9	5	-3.08	-3.34	4.33
	10	4	-1.08	-2.91	3.10
	11	5	-1.93	-3.44	3.94
グループ3	12	4	2.37	-4.13	4.76
	13	4	3.52	-4.53	5.74
	14	5	1.52	-4.35	4.89
	15	5	2.87	-5.06	5.72
	16	6	0.67	-5.18	5.22
	17	6	1.83	-5.58	5.87
	18	0	2.30	-0.81	2.44

【0150】データについて判別分析を実施し判別関数値を求めると、表35の右側、判別関数値の欄のようになる。3グループあるので判別関数値は2組得られる。一般に判別関数値は(グループ数-1)組得られる。この2組の判別関数値から次の式に基づいてユークリッド距離が求められ、これを表35のユークリッド距離の欄に示す。

※判別関数値のユークリッド距離 $d (= (f_1^2 + f_2^2)^{1/2})$

判別関数値のユークリッド距離について各グループごとに U_t を求めると、 U_t は次のようになる。グループ1の U_t を表36に示す。

【0151】

※ 【表36】

グループ1

外れ値	大きい方	なし	4.88
小さい方	なし	0.02	1.41
	1.97	0.34	1.84

【0152】 U_t は、外れ値の候補がない場合が最小である。従って外れ値はない。グループ2の U_t を表37に示す。

★ 【0153】

【表37】

★

グループ2

外れ値	大きい方	なし	4.98
小さい方	なし	-2.92	-2.01
	3.10	-1.16	-0.22

【0154】 U_t は、外れ値の候補がない場合が最小である。従って外れ値はない。グループ3の U_t を表38に示す。

【0155】

【表38】

外れ値 小さい方	大きい方		
	なし	5.87	5.87
なし	0.67	2.65	4.62
2.44	-3.10	-1.49	2.03
2.44, 4.76	-0.94	1.88	4.77

【0156】従ってユークリッド距離が2.44、サンプルno. 18を外れ値とする。なお、この実施例では特性値が3グループの場合について述べたが、3グループ以上についても同様に行える。

【0157】実施例17. この実施例では、上記実施例で示した数2とは異なる計算式で検出統計量を求める場合について述べる。検出統計量Utaの計算式は数3を用いる。

【0158】

【数3】

$$Uta = n \log \alpha + \frac{b_2}{2} + 2S$$

20

ただし n はサンプル数

S は外れ値の候補の個数

α は x_1, \dots, x_n をサンプルデータとすると

$$\alpha^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

$$b_2 = \frac{\sum (x_i - \bar{x})^4}{\alpha^2}$$

30

*

除いたサンプル	$n \log \alpha + b_2 / 2 + 2s$	参考 $n \log \alpha + 2s$
なし	-7.287	-9.419
1.01	-5.860	-8.323
-1.40	-9.770	-11.153 (1)
1.01, -1.40	-10.101 (1)	-11.117
1.01, 0.63	-3.577	-6.237
-1.40, -0.44	-7.421	-8.815
-1.40, -0.44, 1.01	-7.825	-8.806
-1.40, 1.01, 0.63	-8.575	-9.561

【0161】従って、表より1.01、-1.40を外れ値とする。表39の参考の欄を見るとわかるように、 $U = n \log \alpha + 2s$ を用いると-1.40を外れ値としているので、Utaでは1個多く外れ値を指定している。ところが、他の多くのデータでは、数2の計算式を用いた場合と同様の結果を得ている。このことより、数3の計算式は、場合によっては1個多く外れ値を検出するという特徴がある。尚、数3の第2項の $b_2 / 2$ は、「竹内」の正規分布のあてはまりのよさの補正の指標を参考とした。

【0159】データとして(総データ数15)、

-1.40、-0.44、-0.30、-0.24、-0.22、-0.13、-0.15、0.06、0.10、0.18、0.20、0.39、0.48、0.63、1.01

を用いる。計算結果は表39のようになる。

【0160】

【表39】

竹内(竹内 啓(1976): "情報統計量の分布とモデルの適切さの基準"、"数理科学"、NO. 153サイエンス社、12-18)によれば正規分布モデルの適切さを表す統計量(以下竹内の統計量)T_iは、 z_i ($i = 1, \dots, n$)をサンプル数nのデータ、 \bar{z} を z_i の平均として次のようになる。

$$T_i = -\log \sigma - b_2 / 2n$$

ここで、

$$\sigma^2 = \{ \sum (z_i - \bar{z})^2 \} / n$$

$$b_2 = \{ \sum (z_i - \bar{z})^4 \} / n \sigma^4$$

この竹内の統計量の値が大きいかほど適切な正規分布モデルに近い。

【0162】正規分布モデルaと正規分布モデルbの2つがあり、正規分布モデルaの分散は正規分布モデルbの分散よりも小さな値を示す場合、正規分布モデルaの方が正規分布モデルbよりもデータ z_i が平均=0に近い値を多く示す。上記竹内の統計量を求める式の第2項にある $b_2/2n$ は、補正項と呼ばれているものであり、第1項にある $\log \sigma$ の値を補正する意味を持っているものである。従って、竹内の統計量は第1項にある $\log \sigma$ の値が大きく影響するものである。従って、分散 σ の値によってこの竹内の統計量の特徴付けがなされる。従って、分散が小さいほど竹内の統計量の値が大きくなり、この竹内の統計量の値が大きいかほど正規分布モデルbよりも正規分布モデルaに近いパターン（即ち、分散の小さいパターン）を示すことになる。

【0163】実施例18。以下、この実施例18から実施例46までは、検出統計量を求めるための計算式を図10に示す回帰分析説明変数選択基準を基礎にして作成している。前述した数2及び数3はAICを基礎にして考えたものである。AICは、回帰分析説明変数選択基準の一例である。従って、以下の実施例18から実施例46までは、AICによる回帰分析説明変数選択基準以外の回帰分析説明変数選択基準を基礎にして、検出統計量を求める場合においても前述した実施例と同様な効果を奏することができる点について説明する。実施例18から実施例46までに示す検出統計量の計算式数4から*

外れ値の大きい方 検出 小さい方	なし	なし	1.01	1.01 0.63
なし	0.534	0.565	0.667	
-1.40	0.482	0.458	0.537	
-1.40 -0.44	0.547	0.572	0.752	

【0168】従って、-1.40と1.01を外れ値とする。

【0169】実施例19。検出統計量Ftの計算式を、数5に示す。

【0170】

【数5】

*数32は、図10に示す回帰分析説明変数選択基準を基礎にして考えられたものであり、これらの計算式は、大きく分けて2つのタイプに分類できる。第1のグループは、前述した実施例までと同じ形式で、

$n \log \sigma$ + 第2項 (+第3項)

である。第2のグループは、乗算タイプで、

調整因子 $\times \sigma$

である。また、データとして、

-1.40, -0.44, -0.30, -0.24, -0.22, -0.15, -0.13, 0.06, 0.10, 0.18, 0.20, 0.39, 0.48, 0.63, 1.01

を、この実施例以後の全ての実施例で用いる。これを図示すると図11のようになる。同じデータに対して、計算式の違いにより求まる外れ値の数が異なっている。よって、外れ値を多く出さなくてもよい適用業務と、外れ値を多く出したい業務により、計算式を選んで使うことができる。

【0164】検出統計量Stの計算式を、数4に示す。

【0165】

【数4】

$$S_t = \frac{(n-2)(n-1)}{(n-5-2)(n-5-1)} \alpha$$

【0166】Stを計算すると表40のようになる。

【0167】

【表40】

$$F_t = \frac{n+S}{n-5} \alpha$$

【0171】Ftを計算すると表41のようになる。

40 【0172】

【表41】

外れ値の 大きい方 候補 小さい方	なし	1.01	1.01 0.63
なし	0.53	0.55	0.62
-1.40	0.45	0.43	0.46
-1.40 -0.44	0.51	0.49	0.54

【0173】従って、-1.40と1.01を外れ値と 10*
する。

$$T_t = \frac{n^2 - n - S - 2}{n(n - S - 2)(n - S - 1)} \alpha$$

【0174】実施例20. 検出統計量 T_t の計算式を、
数6に示す。

【0176】 T_t を計算すると表42のようになる。

【0175】

【0177】

【数6】

*

【表42】

外れ値の 大きい方 候補 小さい方	なし	1.01	1.01 0.63
なし	0.041	0.046	0.059
-1.40	0.038	0.041	0.052
-1.40 -0.44	0.048	0.055	0.079

【0178】従って、-1.40を外れ値とする。

※

$$TIt = n \log \alpha + 2S + \frac{S(S-2)}{n}$$

【0179】実施例21. 検出統計量 TIt の計算式
を、数7に示す。

【0181】 TIt を計算すると表43のようになる。

【0180】

【0182】

【数7】

※30

【表43】

外れ値の 大きい方 候補 小さい方	なし	1.01	1.01 0.63
なし	-9.42	-8.39	-6.24
-1.40	-11.22	-11.12	-9.31
-1.40 -0.44	-8.82	-8.56	-6.49

【0183】従って、-1.40を外れ値とする。

$$Wt = S \log n + n \log \alpha$$

【0184】実施例22. 検出統計量 Wt の計算式を、
数8示す。

【0186】 Wt を計算すると表44のようになる。

【0185】

【0187】

【数8】

【表44】

41

42

外れ値の 大きい方 候補 小さい方	なし	1.01	1.01 0.63
なし	-9.42	-7.78	-5.11
-1.40	-10.51	-9.99	-8.11
-1.40 -0.44	-7.69	-7.35	-5.83

【0188】従って、-1.40を外れ値とする。

* 【0191】Ptを計算すると表45のようになる。

【0189】実施例23. 検出統計量Ptの計算式を、

【0192】

数9示す。

【表45】

【0190】

【数9】

$$P_t = \frac{1}{(n-S-1)^2} \alpha$$

*

外れ値の 大きい方 候補 小さい方	なし	1.01	1.01 0.63
なし	0.002722	0.0033	0.0045
-1.40	0.002714	0.0031	0.0043
-1.40 -0.44	0.00373	0.0045	0.0069

【0193】従って、-1.40を外れ値とする。

※

$$H_t = \frac{(n-1)(n+S+1)}{(n+1)(n-S-2)} \alpha$$

【0194】実施例24. 検出統計量Ht計算式を、数

30

【0196】Htを計算すると表46のようになる。

10示す。

【0197】

【0195】

※

【表46】

【数10】

外れ値の 大きい方 候補 小さい方	なし	1.01	1.01 0.63
なし	147.14	135.69	135.89
-1.40	110.85	93.36	89.37
-1.40 -0.44	111.45	85.18	96.28

【0198】従って、-1.40、1.01、0.63
を外れ値とする。

$$U_{1t} = n \log \alpha + 2 * S * 1.1$$

【0199】実施例25. 検出統計量U_{1t}の計算式
を、数11示す。【0201】U_{1t}を計算すると表47のようにな
る。

【0200】

【0202】

【数11】

【表47】

外れ値の 大きい方 候補 小さい方	なし	1.01	1.01 0.63
なし	-9.42	-8.32	-6.24
-1.40	-11.15	-11.12	-9.56
-1.40 -0.44	-8.82	-8.81	-7.22

【0203】従って、-1.40を外れ値とする。

10* 【0206】 U_t を計算すると表48のようにな

【0204】実施例26. 検出統計量 U_t の計算式
を、数12示す。

【0207】

【表48】

【0205】

【数12】

$$U_t = n \log \alpha + 2 * S * 0.9$$

*

外れ値の 大きい方 候補 小さい方	なし	1.01	1.01 0.63
なし	-9.42	-8.52	-6.64
-1.40	-11.35	-11.52	-10.16
-1.40 -0.44	-9.22	-9.41	-8.02

【0208】従って、-1.40と1.01を外れ値と
する。

※

$$U_t = \frac{n-1}{n-S} \alpha$$

【0209】実施例27. 検出統計量 U_t の計算式
を、数13示す。

【0211】 U_t を計算すると表49のようになる。

【0212】

【表49】

【0210】

30

※

【数13】

外れ値の 大きい方 候補 小さい方	なし	1.01	1.01 0.63
なし	0.498	0.478	0.498
-1.40	0.391	0.341	0.334
-1.40 -0.44	0.407	0.358	0.358

【0213】従って、-1.40、1.01、0.63
を外れ値とする。

$$B_t = \frac{n+1}{n-S-1} \alpha$$

【0214】実施例28. 検出統計量 B_t の計算式を、
数14示す。

【0216】 B_t を計算すると表50のようになる。

【0217】

【表50】

【0215】

【数14】

45

46

外れ値の 大きい方 候補 小さい方	なし	1.01	1.01 0.63
なし	0.810	0.598	0.637
-1.40	0.489	0.438	0.444
-1.40,-0.44	0.522	0.473	0.501

【0218】従って、-1.40と1.01を外れ値とする。

10

$$Dt = \frac{(n+1)(n+s+1)}{n-s} \alpha$$

【0219】実施例29. 検出統計量Dtの計算式を、数15示す。

【0221】Dtを計算すると表51のようになる。

【0222】

【表51】

【0220】

【数15】

*

外れ値の 大きい方 候補 小さい方	なし	1.01	1.01 0.63	1.01 0.63 0.48
なし	8.539	8.243	8.504	9.350
-1.40	6.734	5.911	5.833	6.063
-1.40,-0.44	7.356	6.212	6.304	6.954

【0223】従って、-1.40、1.01、0.63を外れ値とする。

※

$$Gt = \frac{n+s+1}{n-s+1} \alpha$$

【0224】実施例30. 検出統計量Gtの計算式を、数16示す。

【0226】Gtを計算すると表52のようになる。

【0227】

【表52】

【0225】

【数16】

※

外れ値の 大きい方 候補 小さい方	なし	1.01	1.01 0.63	1.01 0.63 0.48
なし	0.534	0.547	0.807	0.702
-1.40	0.447	0.417	0.438	0.482
-1.40,-0.44	0.498	0.466	0.501	0.583

【0228】従って、-1.40と1.01を外れ値とする。

$$Qt = \frac{1}{(n-s)^2} \alpha$$

【0229】実施例31. 検出統計量Qtの計算式を、数17示す。

40

【0231】Qtを計算すると表53のようになる。

【0232】

【表53】

【0230】

【数17】

47

48

外れ値の 大きい方 候補 小さい方	なし	1.01	1.01 0.63	1.01 0.63 0.48
なし	0.00237	0.0028	0.0038	0.0054
-1.40	0.00231	0.00258	0.0034	0.0049
-1.40,-0.44	0.0031	0.00359	0.0051	0.0097

【0233】従って、-1.40を外れ値とする。

10*

$$I_t = \frac{n+S}{(n-S)^2} \propto$$

【0234】実施例32. 検出統計量 I_t の計算式を、【0236】 I_t を計算すると表54のようになる。

数18示す。

【0237】

【0235】

*

【表54】

【数18】

外れ値の 大きい方 候補 小さい方	なし	1.01	1.01 0.63	1.01 0.63 0.48
なし	0.036	0.043	0.058	0.081
-1.40	0.035	0.039	0.051	0.074
-1.40,-0.44	0.048	0.054	0.077	0.130

【0238】従って、-1.40を外れ値とする。

※【0240】

【0239】実施例33. 検出統計量 V_t の計算式を、

【数19】

数19示す。

※

$$V_t = \frac{n-1}{(n-S)^2} \propto$$

【0241】 V_t を計算すると表55のようになる。

★【表55】

【0242】

★

外れ値の 大きい方 候補 小さい方	なし	1.01	1.01 0.63	1.01 0.63 0.48
なし	0.033	0.037	0.045	0.063
-1.40	0.030	0.031	0.037	0.049
-1.40,-0.44	0.037	0.040	0.051	0.078

【0243】従って、-1.40を外れ値とする。

【0244】実施例34. 検出統計量 E_t の計算式を、

40

$$E_t = \frac{n+1}{(n-S-1)^2} \propto$$

数20示す。

【0246】 E_t を計算すると表56のようになる。

【0245】

【0247】

【数20】

【表56】

49

50

外れ値の 大きい方 候補 小さい方	なし	1.01	1.01 0.63	1.01 0.63 0.48
なし	0.038	0.043	0.055	0.075
-1.40	0.035	0.038	0.047	0.087
-1.40,-0.44	0.045	0.050	0.070	0.122

【0248】従って、-1.40を外れ値とする。

*

$$J_t = \frac{n(n+3+1)}{(n-5)^2} \alpha$$

【0249】実施例35. 検出統計量Jtの計算式を、

10

【0251】Jtを計算すると表57のようになる。

数21示す。

【0252】

【0250】

*

【表57】

【数21】

外れ値の 大きい方 候補 小さい方	なし	1.01	1.01 0.63	1.01 0.63 0.48
なし	0.569	0.634	0.782	1.040
-1.40	0.518	0.537	0.648	0.866
-1.40,-0.44	0.641	0.690	0.901	1.391

【0253】従って、-1.40を外れ値とする。

【0254】実施例36. 検出統計量Vtdの計算式

を、数22示す。

【0255】

【数22】

$$Utd = n \log \alpha - \frac{b_2}{2} + 2 * S$$

【0256】Vtdを計算すると表58のようになる。

【0257】

【表58】

外れ値の 大きい方 候補 小さい方	なし	1.01	1.01 0.63
なし	-11.57	-10.79	-8.90
-1.40	-12.54	-12.13	-10.55
-1.40,-0.44	-10.21	-9.79	-8.15

【0258】従って、-1.40を外れ値とする。

【0259】実施例37. 検出統計量BBtの計算式

を、数23示す。

【0260】

【数23】

$$BBt = \frac{n+1}{(n-5-1)^2} \alpha$$

【0261】BBtを計算すると表59のようになる。

【0262】

【表59】

外れ値の 大きい方 候補 小さい方	なし	1.01	1.01 0.63
なし	0.044	0.050	0.064
-1.40	0.041	0.044	0.056
-1.40,-0.44	0.052	0.059	0.084

【0263】従って、-1.40を外れ値とする。

【0264】実施例38. 検出統計量CCtの計算式を、数24示す。

【0265】

【数24】

$$CCt = \frac{n(n-5+1)}{(n-5)^2} \alpha$$

【0266】CCtを計算すると表60のようになる。

【0267】

【表60】

外れ値の 大きい方 候補 小さい方	なし	1.01	1.01 0.63
なし	0.569	0.559	0.587
-1.40	0.453	0.403	0.405
-1.40,-0.44	0.481	0.431	0.450

【0268】従って、-1.40と1.01を外れ値とする。

50

【0269】実施例39. 検出統計量DDtの計算式を、数25示す。

【0270】

【数25】

$$DDt = \frac{n(n+S+1)}{(n-S)^2} \sim$$

【0271】DDtを計算すると表61のようになる。

【0272】

【表61】

外れ値の 大きい方 候補 小さい方	なし	1.01	1.01 0.63
なし	0.589	0.634	0.782
-1.40	0.518	0.537	0.648
-1.40, -0.44	0.641	0.690	0.901

【0273】従って、-1.40を外れ値とする。

【0274】実施例40. 検出統計量GGtの計算式を、数26示す。

【0275】

【数26】

$$GGt = \frac{n+S+1}{(n-S+1)^2} \sim$$

【0276】GGtを計算すると表62のようになる。

【0277】

【表62】

外れ値の 大きい方 候補 小さい方	なし	1.01	1.01 0.63
なし	0.033	0.039	0.051
-1.40	0.032	0.035	0.044
-1.40, -0.44	0.042	0.047	0.063

* 【0278】従って、-1.40を外れ値とする。

【0279】実施例41. 検出統計量Up tの計算式を、数27示す。

【0280】

【数27】

$$Up t = \frac{n+S-1}{(n-S+1)^2} \sim$$

【0281】Up tを計算すると表63のようになる。

【0282】

【表63】

外れ値の 大きい方 候補 小さい方	なし	1.01	1.01 0.63
なし	0.487	0.478	0.531
-1.40	0.391	0.365	0.383
-1.40, -0.44	0.435	0.408	0.439

10

【0283】従って、-1.40と1.01を外れ値とする。

【0284】実施例42. 検出統計量Ztの計算式を、数28示す。

【0285】

【数28】

$$Zt = \frac{n}{n-S} \sim$$

【0286】Ztを計算すると表64のようになる。

【0287】

【表64】

30

*

外れ値の 大きい方 候補 小さい方	なし	1.01	1.01 0.63	1.01 0.63 0.48
なし	0.534	0.515	0.538	0.585
-1.40	0.421	0.369	0.365	0.379
-1.40, -0.44	0.441	0.388	0.394	0.435

【0288】従って、-1.40、1.01、0.63を外れ値とする。Ztは外れ値を多めに検出する。

【0289】実施例43. 検出統計量Ktの計算式を、数29示す。

【0290】

【数29】

$$Kt = \frac{1}{n-S} \sim$$

【0291】Ktを計算すると表65のようになる。

【0292】

【表65】

50

53

54

外れ値の 大きい方 候補 小さい方	なし	1.01	1.01 0.63	1.01 0.63 0.48
なし	0.036	0.037	0.041	0.049
-1.40	0.030	0.028	0.030	0.034
-1.40, -0.44	0.034	0.032	0.038	0.043

【0293】従って、-1.40と1.01を外れ値とする。

【0294】実施例44. 検出統計量 X_t の計算式を、数30示す。

【0295】

【数30】

$$X_t = \frac{1}{(n-S)(n+S)} \alpha$$

【0296】 X_t を計算すると表66のようになる。

【0297】

【表66】

外れ値の 大きい方 候補 小さい方	なし	1.01	1.01 0.63
なし	0.0024	0.0025	0.0028
-1.40	0.0020	0.0019	0.0020
-1.40, -0.44	0.0023	0.0022	0.0024

【0298】従って、-1.40と1.01を外れ値とする。

【0299】実施例45. 検出統計量 HQ_t の計算式を、数31示す。

【0300】

【数31】

$$HQ_t = n \log \alpha + C * S * \log (\log (n))$$

$$C > 2$$

【0301】 HQ_t を計算すると表67のようになる。

【0302】

【表67】

外れ値の 大きい方 候補 小さい方	なし	1.01	1.01 0.63
なし	-9.42	-7.41	-4.59
-1.40	-10.24	-9.47	-7.40
-1.40, -0.44	-7.16	-6.61	-4.72

【0303】従って、-1.40を外れ値とする。

10 【0304】実施例46. 検出統計量 AIC_t の計算式を、数32示す。

【0305】

【数32】

$$AIC_t = n \log \alpha^2 + 2 * S$$

【0306】 AIC_t を計算すると表68のようになる。

【0307】

【表68】

20

外れ値の 大きい方 候補 小さい方	なし	1.01	1.01 0.63
なし	-17.80	-17.61	-15.43
-1.40	-23.27	-25.19	-24.08
-1.40, -0.44	-20.59	-22.57	-21.39

【0308】従って、-1.40と1.01を外れ値とする。

【0309】

【発明の効果】第1の発明によれば、値を入力すれば算出された検出統計量に基づき外れ値が検出されるので、従来のように計算値と数表の大小比較をする必要がない。また、外れ値の個数を予め設定する必要がない。あるいは、外れ値として検出したい数の最大値を指定しておけばよい。また、外れ値の個数により、又は大きい方の外れ値が、小さい方の外れ値かにより計算方式を変える必要もないので、計算過程が簡単でかつ計算量も少なくすむ。また、従来方式では、マスク効果により外れ値を検出できないことがあったが、これを回避することができるので、より正確な結果が得られる。また、外れ値が存在しない時は、存在しないと判定する。

【0310】第2の発明によれば、外れ値を算出するための値の選択が簡単に行える。

【0311】第3の発明によれば、検出統計量の単純な比較だけで外れ値を求めるので、処理が簡単になる。

【0312】第4の発明における計算式によれば、外れ値がある場合、最も外れた値が除かれると検出統計量は

最小となるので、これを利用して外れ値を求めることができる。

【0313】第5の発明における計算式によれば、補正項があるため、よりの確に外れ値を求めることができる。

【0314】第6の発明における計算式によれば、第1の項目に分散を含んでいるため、最も外れている値を除くと分散が小さくなるという性質を利用できる。

【0315】第7の発明における計算式によれば、第2の項目に外れ値の候補の個数を含んでいるため、外れ値の個数を増やしていったことによる第1の項目の減少傾向を相殺できる。

【0316】第8の発明における計算式によれば、第2の項目に係数を乗算していることにより、第2の項目の増加量を調節することができる。

【0317】第9の発明における計算式によれば、分散の減少傾向に係数により、調節できる。

【0318】第10の発明における計算式によれば、回帰分析の式を応用して外れ値の検出を行うことができる。

【0319】第11の発明によれば、加工工程があることによりさまざまなタイプのデータの外れ値を検出することができる。

【0320】第12の発明によれば、時間に依存する値からも時間に依存しない値に加工することにより、外れ値を検出することができる。

【0321】第13の発明によれば、1つのサンプルに複数の特性値が存在する場合であっても、デコ比を計算することにより外れ値を求めることができる。

【0322】第14の発明によれば、回帰分析の手法を適用できる値であれば、回帰分析の残差を求めることによりこの残差の外れ値を求めることができる。

【0323】第15の発明によれば、正準相関分析モデルを適用できる場合でも、外れ値を求めることができる。

【0324】第16の発明によれば、複数のグループに特性値が分類され、判別分析を行うことができる場合、外れ値を求めることができる。

【0325】第17の発明によれば、上記のような外れ値検出方法を利用することにより、外れ値を容易に検出することができるデータ処理装置を得ることができる。この外れ値が得られた時の環境条件を検討することによ

り、新たな知見・情報が得られることにもなる。

【0326】第18の発明によれば、上記のような外れ値検出方法を利用することにより、外れ値を容易に検出し、除くことができるデータ処理装置を得ることができる。このデータ処理装置により得られた結果は、外れ値を除いてあるので信頼性が向上している。

【図面の簡単な説明】

【図1】本発明の外れ値検出方法を説明するためのフローチャート図である。

【図2】Grubbsのデータ1を用いた場合の外れ値の候補の個数 s と検出統計量 U_t の関係を示す図である。

【図3】別のデータを用いた場合の外れ値の候補の個数 s と検出統計量 U_t の関係を示す図である。

【図4】本発明のデータ処理装置の構成図である。

【図5】本発明の一実施例の入力データをプロットした図である。

【図6】本発明の外れ値検出のための工程を説明する図である。

【図7】本発明の一実施例の時間とともに増加する傾向を持つデータをプロットした図である。

【図8】本発明の一実施例の時間とともに増加する傾向を除いたデータをプロットした図である。

【図9】本発明の一実施例の入力データをプロットした図である。

【図10】回帰分析説明変数選択基準の数式を示す図である。

【図11】本発明の実施例の中で使われるデータをプロットした図である。

【図12】従来の技術及び本発明の実施例の中で使われるデータをプロットした図である。

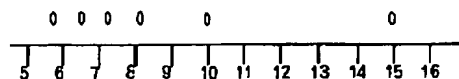
【図13】従来の技術及び本発明の実施例で使われる装置の構成図である。

【図14】従来の外れ値検出方式を説明するためのフローチャート図である。

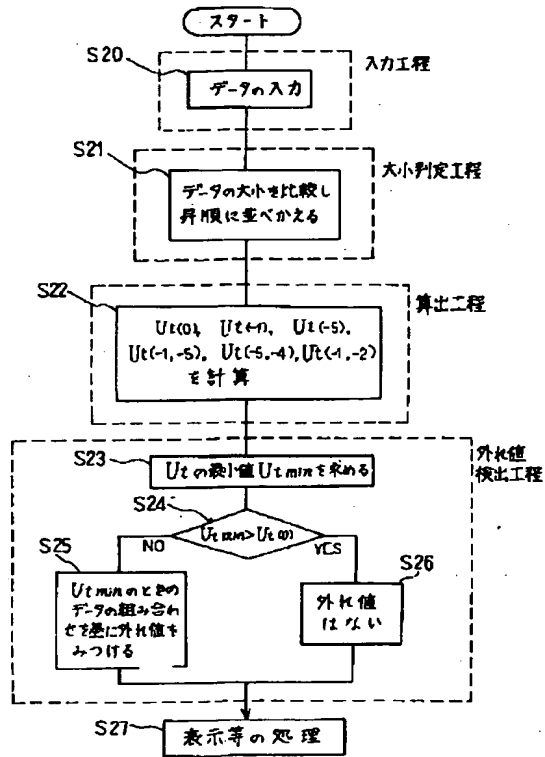
【符号の説明】

- 1 情報処理装置
- 2 コンピュータ (FDD付き)
- 3 ディスプレイ・ユニット
- 4 プリンタ
- 5 キーボード
- 6 フロッピーディスク

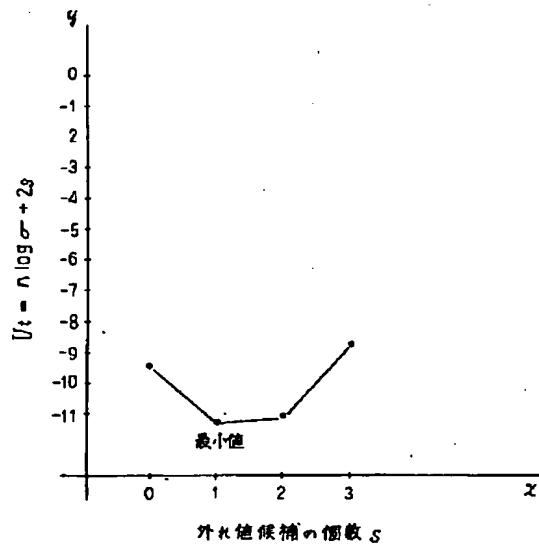
【図12】



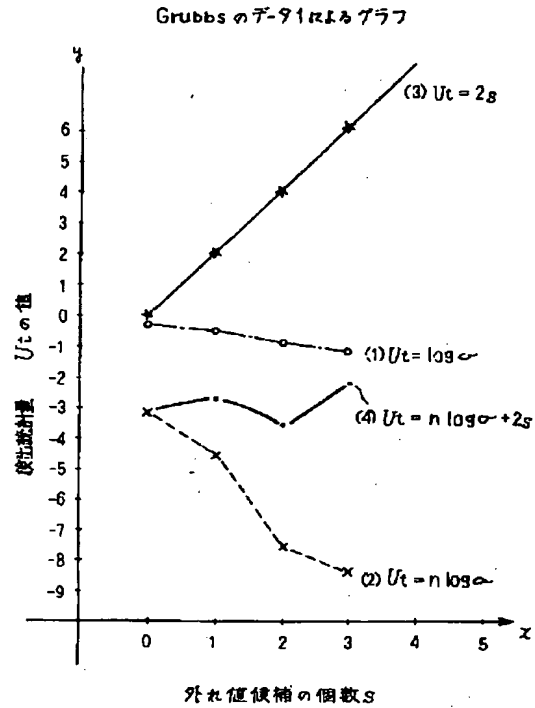
【図1】



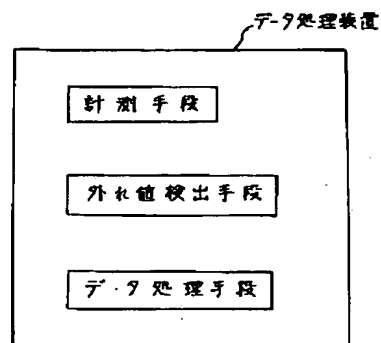
【図3】



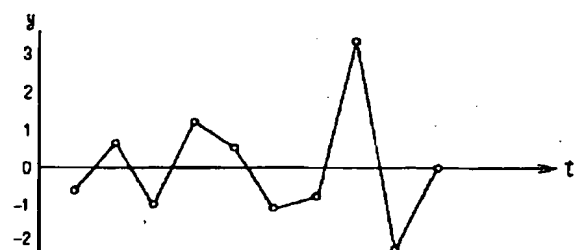
【図2】



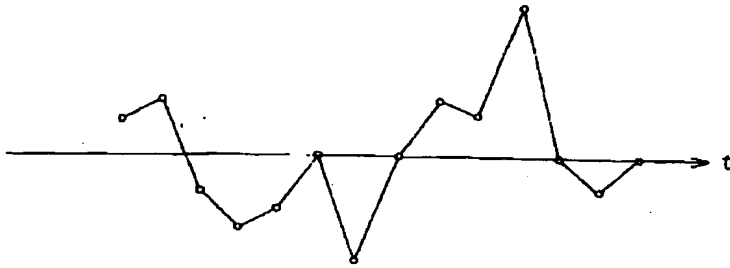
【図4】



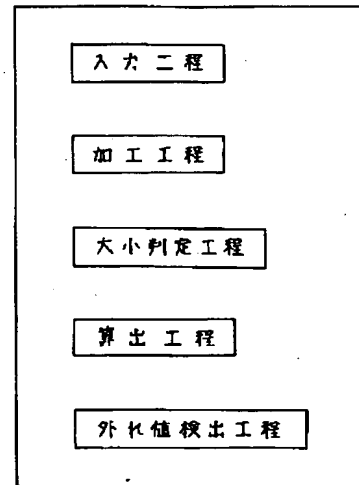
【図8】



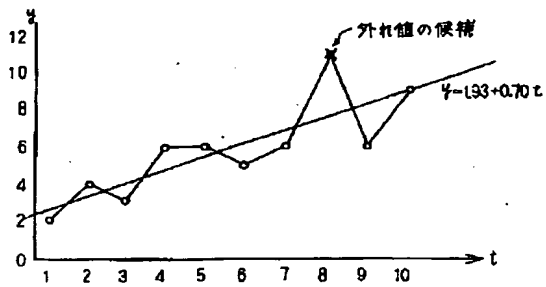
【図5】



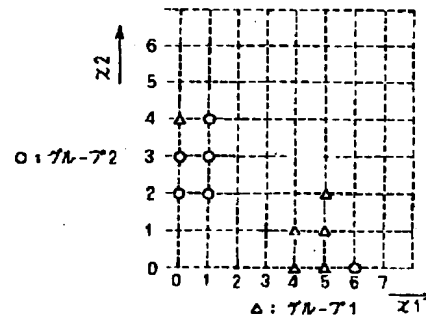
【図6】



【図7】

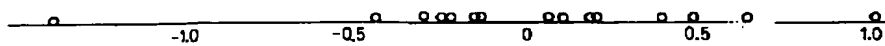


【図9】



【図11】

実施例18~46のデータのプロット

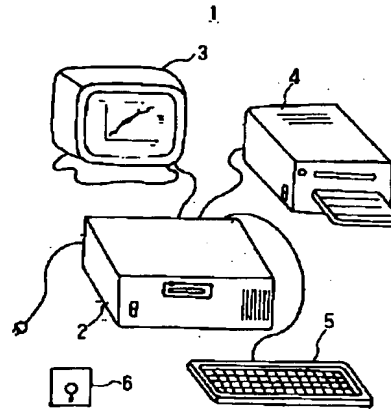


【図10】

回帰分析説明変数選択規準

- 1 $aic = n \log (\sum ci^2) + 2P$ n : データの個数
- 2 $aic = aic + P(P-2)/n$ P : 説明変数の個数
- 3 $校正 = 1 - \frac{(n-2)(n-1)}{(n-P-2)(n-P-1)} (1-R^2)$ R : 重相関係数
- 4 $調整 = 1 - \frac{(n-1)(n+P+1)}{(n+1)(n-P-1)} (1-R^2)$
- 5 $within = \frac{n^2 - n - P - 2}{n(n-P-2)(n-P-1)} \sigma^2$ σ^2 : 残差の分散
- 6 $予測 = \frac{(n+1)(n-2)}{n(n-P-2)} \sigma^2$
- 7 $SE = \frac{\sum ei^2}{(n-P-1)(n-P-2)}$ $\sum ei^2$: 残差の2乗和
- 8 $上相 = 1 - \frac{(n+P+1)}{(n-P-1)} (1-R^2)$
- 9 $aicc = n \log (\sum ci^2) + C \cdot P$ $C > 0$
- 10 $HQ = n \log (\sum ci^2) + C \cdot P \cdot \log (\log (n))$ $C > 2$
- 11 $fpe = \frac{n+P}{n-P} \sum ei^2$
- 12 $Helms = \frac{P}{n(n-P)} \sum ei^2$
- 13 $Shibata = P \log n + n \log \sigma$

【図13】



- 1: 情報処理装置
- 2: コンピュータ(FDD付き)
- 3: ディスプレイ
- 4: プリンタ
- 5: キーボード
- 6: フロッピーディスク

【図14】

従来方式

危険率5%(棄却水準5%)の場合で説明する。外れ値の個数が1あるいは2個のときとする。

